Lieff Cabraser Heimann & Bernstein
Attorneys at Law

Susman Godfrey l.l.p.
a registered limited liability partnership

CDAS

March 31, 2025

VIA ECF

Hon. Ona T. Wang
Daniel Patrick Moynihan
United States Courthouse
500 Pearl St.
New York, NY 10007

RE: *Authors Guild v. OpenAI Inc.*, 23-cv-8292 (S.D.N.Y.) and *Alter v. OpenAI Inc.*, 23-cv-10211 (S.D.N.Y.)

Dear Judge Wang:

Pursuant to Rule II(b) of Your Honor's Individual Practices, Plaintiffs seek a conference regarding Microsoft's refusal to produce for inspection data it [redacted] The full contents of this data bear directly on Plaintiffs' direct and contributory infringement claims. Microsoft's arguments against production are contrary to this Court's prior orders and otherwise without merit.

I. BACKGROUND

For much of this litigation, this case has been focused on OpenAI's illegal acquisition and use of the pirated books dataset known as Library Genesis or LibGen. For this reason, OpenAI has willingly produced a copy of LibGen for remote inspection.

Recent document productions from Microsoft, however, have revealed that Microsoft may have separately downloaded and provided a copy of LibGen for training of OpenAI's models. *See* Ex. 1, MSFT_AICPY_000591635 (email from October 1, 2023, 2:40 pm, [redacted] …"); Ex. 2, MSFT_AICPY_000591895 (lists [redacted]); Ex. 3, MSFT_AICPY_000340737 ("[redacted]"); Ex. 4, MSFT_AICPY_000212393 ([redacted]").

[redacted] *See id.* ("[redacted]"); Ex. 5, MSFT_AICPY_000169741 ("[redacted]"); Ex. 6, MSFT_AICPY_00588353 [redacted]

[redacted].”); Ex. 7, MSFT_AICPY_000225215 (email from OpenAI to Microsoft [redacted]”).

Soon after becoming aware of this evidence, Plaintiffs requested that Microsoft produce for inspection a copy of LibGen and any other pirated data that was scraped for use by OpenAI for inspection. This inspection request is covered by, *inter alia*, RFP No. 41 which seeks “Documents sufficient to identify all datasets containing commercial works of fiction and/or nonfiction, including Class Works, that You or OpenAI accessed, downloaded or copied to train Large Language Models.” Ex. 8.[1]

On March 5, Microsoft refused Plaintiffs’ request, explaining that the data identified by Plaintiffs “does not appear to relate to training data used for the relevant models—[Microsoft] does not believe further information in this area is within the scope of discovery.” Ex. 9.

Notwithstanding the above, Microsoft argues that the requested data is not relevant because (1) when Microsoft provided the data in late 2023, OpenAI had already trained the GPT-4 Turbo model and none of the later models are relevant and (2) because OpenAI did not train any models on the data that Microsoft provided. *See* Ex. 10 (March 14 email from Microsoft). The parties conferred on these issues and are at an impasse. Plaintiffs thus move to compel Microsoft to produce for inspection the training data it provided to OpenAI in late 2023.

II. ARGUMENT

A. Microsoft’s Copy of LibGen is Highly Relevant to Plaintiffs’ Claims of Direct and Contributory Infringement.

Microsoft’s copy of LibGen is highly relevant for at least two reasons. *First*, the existence and the contents of such a dataset go to willfulness on Plaintiffs’ direct infringement claim against Microsoft. LibGen is a well-known pirated dataset, and evidence that Microsoft downloaded and saved this dataset is evidence of Microsoft’s willfulness. *See, e.g., FameFlynet, Inc. v. Shoshanna Collection, LLC*, 282 F. Supp. 3d 618, 627 (S.D.N.Y. 2017) (finding defendant liable for willful copyright infringement for knowingly saving the plaintiff’s copyrighted photos from an online website). *Second*, the existence and contents of such a dataset may be further evidence of Microsoft’s contributory infringement. As described above, Microsoft provided data to OpenAI, and Plaintiffs have reason to believe that this data included LibGen. Whether LibGen was included in the data Microsoft gave to OpenAI goes to Microsoft’s knowledge that OpenAI was using LibGen to train its Large Language Models. *See, e.g., Abbey House Media, Inc. v. Apple Inc.*, 66

[1] Additional pertinent requests include RFP No. 17&18 (DOCUMENTS and COMMUNICATIONS CONCERNING or RELATING TO the use of commercial works of fiction or nonfiction to train CHATGPT.); No. 23 (“DOCUMENTS and COMMUNICATIONS reflecting or discussing how YOU or OPENAI accessed any commercial works of fiction or nonfiction, INCLUDING FICTION CLASS WORKS AND NONFICTION CLASS WORKS, used to train CHATGPT.”); No. 26 (“DOCUMENTS and COMMUNICATIONS between YOU and OPENAI related to the contents of the training dataset used to train CHATGPT.”); No. 33 (“DOCUMENTS and COMMUNICATIONS exchanged between YOU and OPENAI CONCERNING the use of purportedly copyrighted material in training CHATGPT.”).

F. Supp. 3d 413, 419 (S.D.N.Y. 2014) (explaining that persons who "know or have reason to know of the direct infringement" can be liable for contributory infringement).

B. Microsoft's Arguments Fail.

First, Microsoft's argument that the data it gave OpenAI in late 2023 is not relevant lacks merit. The Court has already ruled that the relevant models extend beyond GPT-4 Turbo and include later models like GPT-4o (released in May 2024) and GPT-4o Mini (released in July 2024). *See* Dkt. No. 293 (limiting the scope of discovery to models set forth in OpenAI's response to Interrogatory No. 11); *see also* Ex. 11 (OpenAI's Supplemental Response to Interrogatory No. 11, listing [redacted]). Moreover, the models are covered by the operative complaint, which alleges that the infringing activity is ongoing and seeks an injunction to stop the activity. *See* Dkt. 69 at ¶¶76, 430.

Second, Microsoft asserts [redacted]. However, [redacted] is beside the point. Plaintiffs are entitled to inspect the data [redacted] constitutes an act of direct infringement. *See, e.g., Am. Geophysical Union v. Texaco Inc.*, 60 F.3d 913, 920 (2d Cir. 1994)*; see also* Ex. 6, MSFT_AICPY_00588353 [redacted]. Under *Texaco*, Plaintiffs can prove direct infringement by showing that Microsoft copied Plaintiffs' works in late 2023. 60 F.3d at 920. Microsoft's scraping of LibGen is evidence of infringement under *Texaco*, and Microsoft must make the data it downloaded in late 2023 available for inspection.

The contents of the data Microsoft sourced for OpenAI bears directly on Plaintiffs' claims. Plaintiffs respectfully request that the Court overrule Microsoft's objections and compel Microsoft to produce for inspection all data it scraped and provided to OpenAI's for use or potential use in its LLMs.

Sincerely,

LIEFF CABRASER HEIMANN & BERNSTEINS LLP	SUSMAN GODFREY LLP	COWAN, DEBAETS, ABRAHAMS & SHEPPARD LLP
/s/ Rachel Geman	/s/ Rohit Nath	/s/ Scott J. Sholder
Rachel Geman	Rohit Nath	Scott J. Sholder